Overview of TREC 2007

Sponsored by:
NIST, IARPA

Ellen Voorhees
NIST
National Institute of Standards and Technology
U.S. Department of Commerce

TREC 2007 Program Committee

Ellen Voorhees, chair
James Allan
Chris Buckley
Gord Cormack
Sue Dumais
Donna Harman
Bill Hersh
David Lewis
John Prager
Steve Robertson
Mark Sanderson
Ian Soboroff
Richard Tong
Ross Wilkinson

TREC 2007 Track Coordinators

- blog: Craig Macdonald,Iadh Ounis,Ian Soboroff
- enterprise: Bailey, Craswell, de Vries, Soboroff
- genomics: Bill Hersh
- legal: Baron, Oard, Thompson, Tomlinson
- million query: James Allan
- QA: Hoa Dang, Diane Kelly, Jimmy Lin
- spam: Gord Cormack

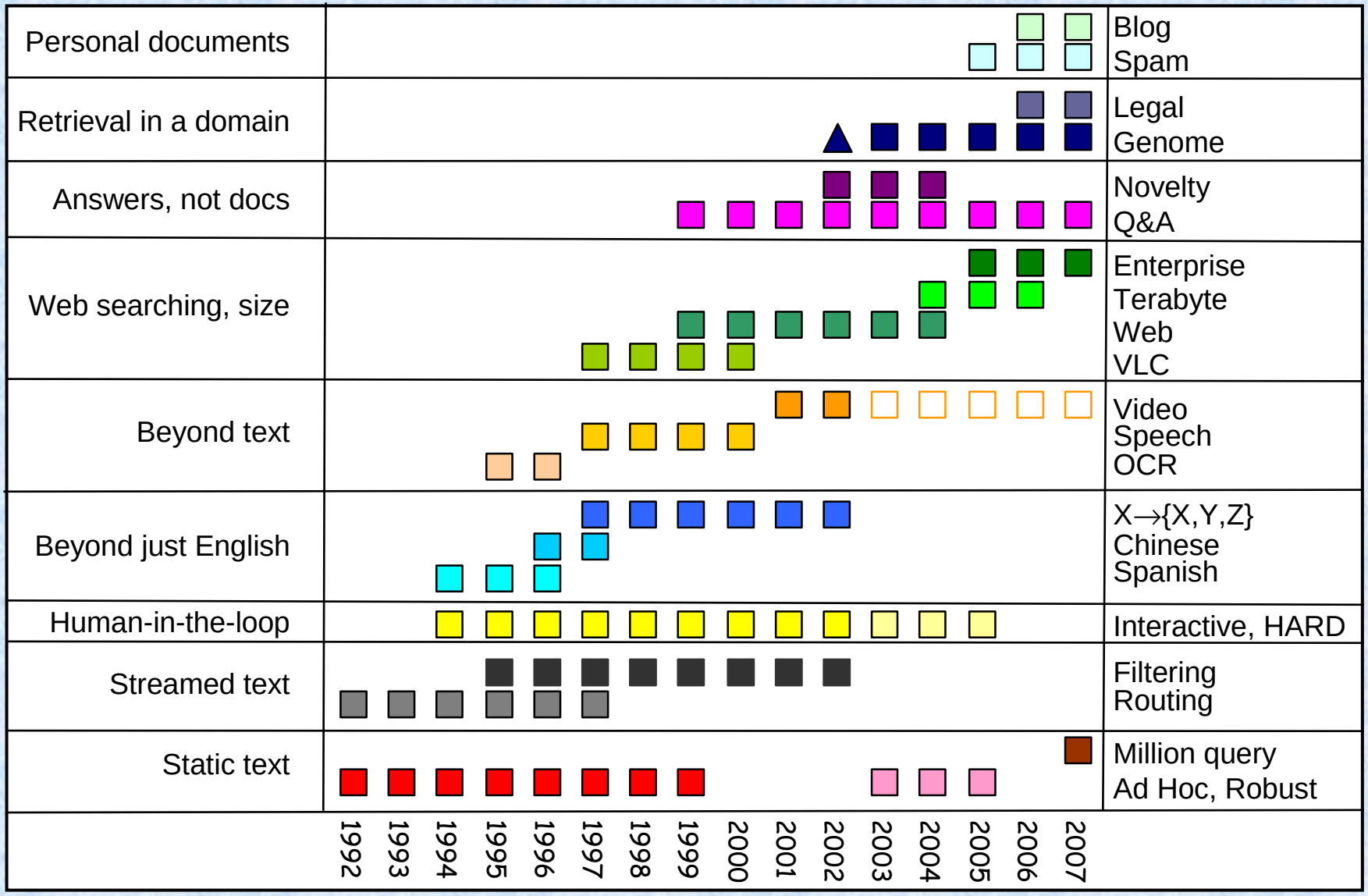
TREC 2007 Participants

Arizona State U.
Beijing U. Posts & Telecom
Carnegie Mellon U.
CMU & U. Karlsruhe
Chinese Acad Sciences (2)
Concordia U. (2)
CRM114
CSIRO ICT Centre
Dalhousie U.
DaLian U. of Technology
Dartmouth College
Drexel U.
EffectiveSoft Ltd.
European Bioinformatics Inst
Exegy Inc.
Fitchburg State College
Fondazione Ugo Bordoni, It. Research Council & U Rome
Fudan U.
Heilongjiang Inst. of Tech.
IBM Cairo
IBM Research, Haifa
Indian Inst. of Tech. Delhi
Illinois Inst. of Tech.
Indiana U.
Int'l. Inst. of Info Tech.
Jozef Stefan Institute
Kobe U. (2)
Kyoto U.
Language Computer Corp
Long Island U.
Lymba Corp.
Massachusetts Inst Tech
Michigan State U.
The MITRE Corp.
National Lib. of Medicine
National Taiwan U.
National U Defense Tech
Northeastern U.
Open Text Corp.
The Open U.
Oregon Health & Sci. U.
Peking U.
Queens College, CUNY
RMIT U.
The Robert Gordon U.
Saarland U.
Sabir Research, Inc.
Shanghai Jiao Tong U (2)
S. China U. of Tech.
St. Petersburg U & INRIA
SUNY Albany
SUNY Buffalo
Technical U. Berlin
TNO, U.Twente & EMC
Tokyo Inst. of Tech.
Tsinghua U.
Tufts U.
Twente U.
U. of Alaska Fairbanks
U. of Alicante
U. of Amsterdam (2)
U. Arkansas Little Rock
U. Colorado Medical Sch
U. of Glasgow
U. & Hospitals of Geneva
U. of Illinois Chicago (2)
U. Illinois Urbana Champaign
U. of Iowa (2)
U. Lethbridge
U. of Maryland
U. of Massachusetts
The U. of Melbourne (2)
U. Missouri Kansas City
U. de Neuchatel
U. North Carolina
U. Rome "La Sapienza"
U. of Strathclyde
U. Texas Austin
U. of Washington
U. of Waterloo (2)
Ursinus College
Weill Cornell Medical Ctr.
Wuhan U.
York U.
Zhejiang U.

TREC Goals

- To increase research in information retrieval based on large-scale collections
- To provide an open forum for exchange of research ideas to increase communication among academia, industry, and government
- To facilitate technology transfer between research labs and commercial products
- To improve evaluation methodologies and measures for information retrieval
- To create a series of test collections covering different aspects of information retrieval

The TREC Tracks



Common Terminology

- "Document" broadly interpreted,
  - for example
    - email message in enterprise, spam tracks
    - blog posting plus comments in blog track
- Tasks
    - ad hoc search: collection known; new queries
    - filtering: standing queries; streaming documents
    - focused response
    - categorization

TREC 2007

- Continue exploring broad themes of 2006
  - only one track change terabyteà million query (similar goals, though different approach)
- Heterogeneous contexts
  - different document genres
    - newswire (QA); web (million query)
    - blogs (blog, QA); email (spam); corporate repositories (legal, enterprise); scientific reports (genomics, legal)
  - different tasks
    - ad hoc, categorization, focused response (QA, passage/entities, experts)

"Million" Query

- Two goals
  - ad hoc retrieval task using large collection (~425GB GOV2 document set; 10000 queries)
  - test specific evaluation hypothesis:
    - a test collection built from very many topics with tens of judgments each is a better diagnostic tool than one built using tens of topics with many judgments each

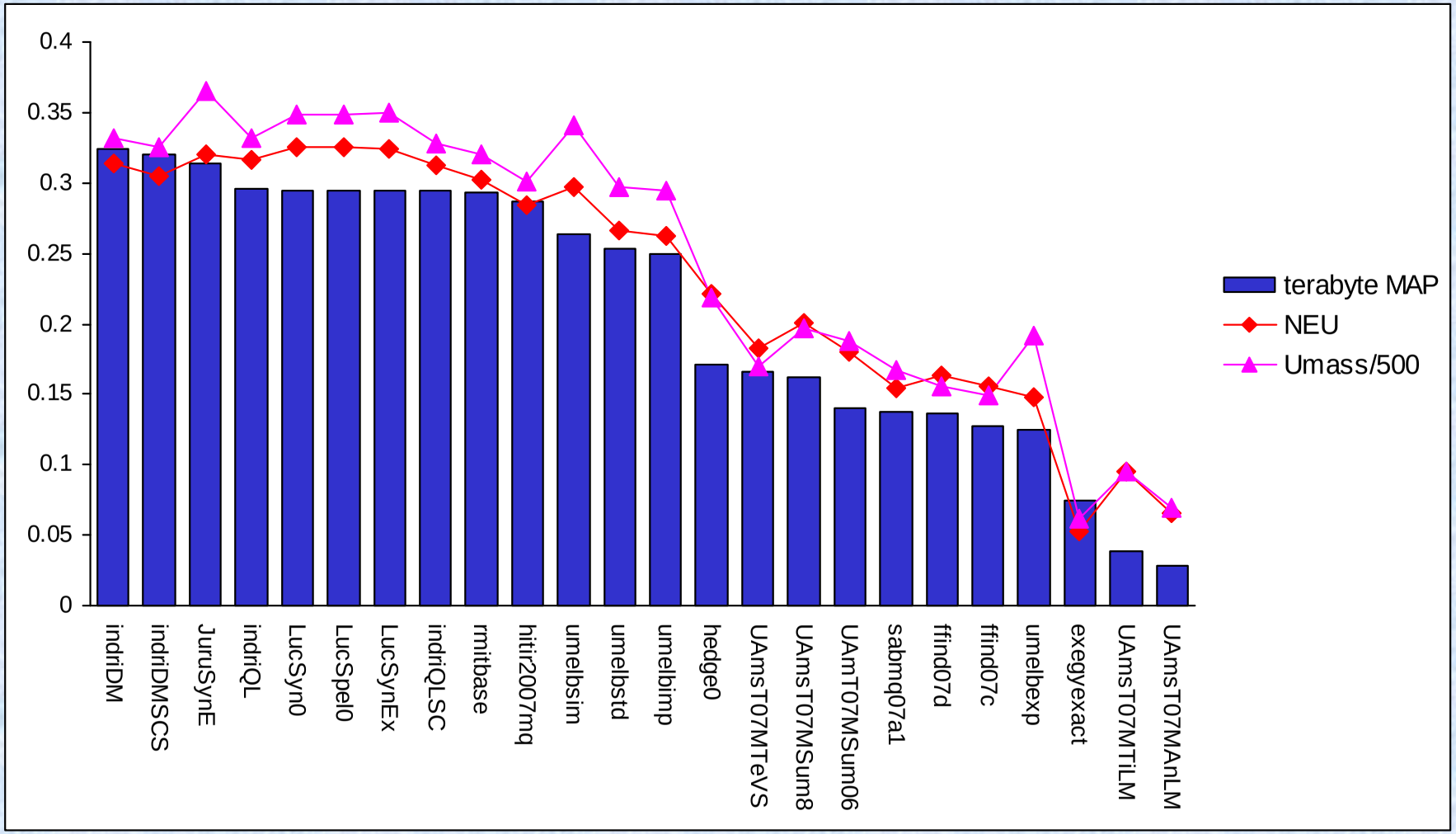
Track Protocol

- Participants run systems against GOV2 using all 10,000 queries
  - queries include previous years' terabyte topics
  - queries taken from web search engine log and had at least one click-through to a GOV2 document
- Some queries judged
  - NIST assessors, participants, others did judging
  - judge selected query from set of 5; made topic statement
  - presented with 40 documents from retrieved docs selected according to particular plan
  - small set of queries judged more than once

Track Protocol (cont'd)

- Sampling method for docs to be judged
  - ¼ queries: 40 docs selected using U. Mass method
  - ¼ queries: 40 docs selected using NEU method
  - ½ queries: 20 docs selected by UMass+20 by NEU
  - actual: 1755 queries judged of which
    - 443 UMass method only & 471 NEU method only
    - 432 alternated with UMass first; 409 with NEU first
- Evaluate runs by
  - average scores as computed by UMass method
  - average scores as computed by NEU method
  - standard trec-eval using terabyte topics and qrels

Relative Effectiveness



Open Issues

- What is Truth?
  - terabyte qrels known to have issues, too
- Reusability?
- How few queries is sufficient?
  - 1755*40 = 70,200 judgments
  - 70,200 judgments ~ 1400/topic for 50 topics

Question Answering Track

- Goal: return answers, not document lists
- Tasks:
  - define a target by answering a series of factoid and list questions about that target, plus returning other info not covered by previous questions
  - complex interactive question answering (ciQA)

Question Series Task

- Same basic task since 2004
  - set of questions to define a target
- Big difference: corpus used as source of answers
  - AQUAINT-2 newswire collection plus
  - blog06 collection
    - much more informal language usage
    - determining "globally correct" answers more difficult
- Scoring change
  - use pyramid-weighted nuggets for Other questions scores

Question Series

254	House of Chanel	
254.1	FACT	Who founded the House of Chanel?
254.2	FACT	In what year was the company founded?
254.3	FACT	Who is the president of the House of Chanel?
254.4	FACT	Who took over the House of Chanel in 1983?
254.5	LIST	What women have worn Chanel clothing to award ceremonies?
254.6	LIST	What museums have displayed Chanel clothing?
254.7	FACT	What Chanel creation is the top-selling fragrance in the world?
254.8	Other	

70 series in test set with 6-10 questions per series

- 19 People
- 17 Organizations
- 19 Things
- 15 Events

- 360 total factoid questions
- 85 total list questions
- 70 total "other" questions

Globally Correct Judgments

- Introduced in 2006
  - need correct time-frame for event targets
  - present tense implied most recent in corpus
- Expanded in 2007
  - a response supported by a document is assumed to be globally correct unless a *better*, *contradictory* answer is supported elsewhere in document collection
    - e.g. nomination of Harriet Miers reported in several newspapers as Oct 3, but as Oct 4 on blog page; blog answer judged locally correct
  - "better", "contradictory" in eyes of the assessor

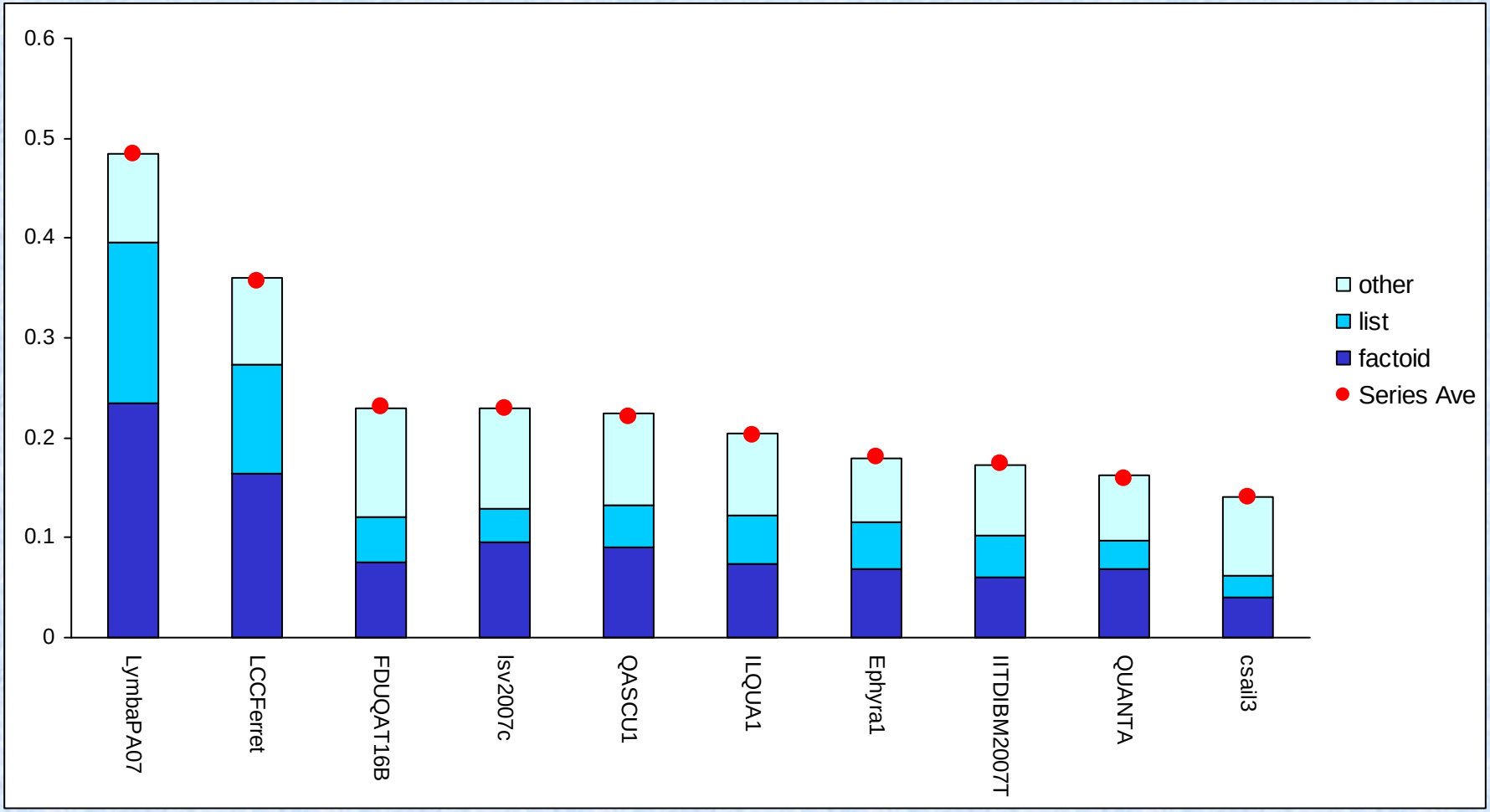
Series Score

- Score a series using weighted average of components

  Score = 1/3FactoidScore+1/3ListScore+1/3OtherScore

- Component score is mean of scores for questions of that type in given series
  - **FactoidScore**: average accuracy. Individual question has score of 1 or 0
  - **ListScore**: average F measure. Recall & precision of response based on set of known answers.
  - **OtherScore**: F(β=3) for that series' Other question, calculated using pyramid-weighted nuggets

Series Task Results



Complex Interactive QA

- Goals:
  - investigate richer user contexts within QA
  - have (limited) actual interaction with user
- Task inspired by TREC 2005 relationship QA task and HARD track
  - "essay" questions
  - interaction forms allowed participants to solicit information from assessor (surrogate user)

Complex Questions

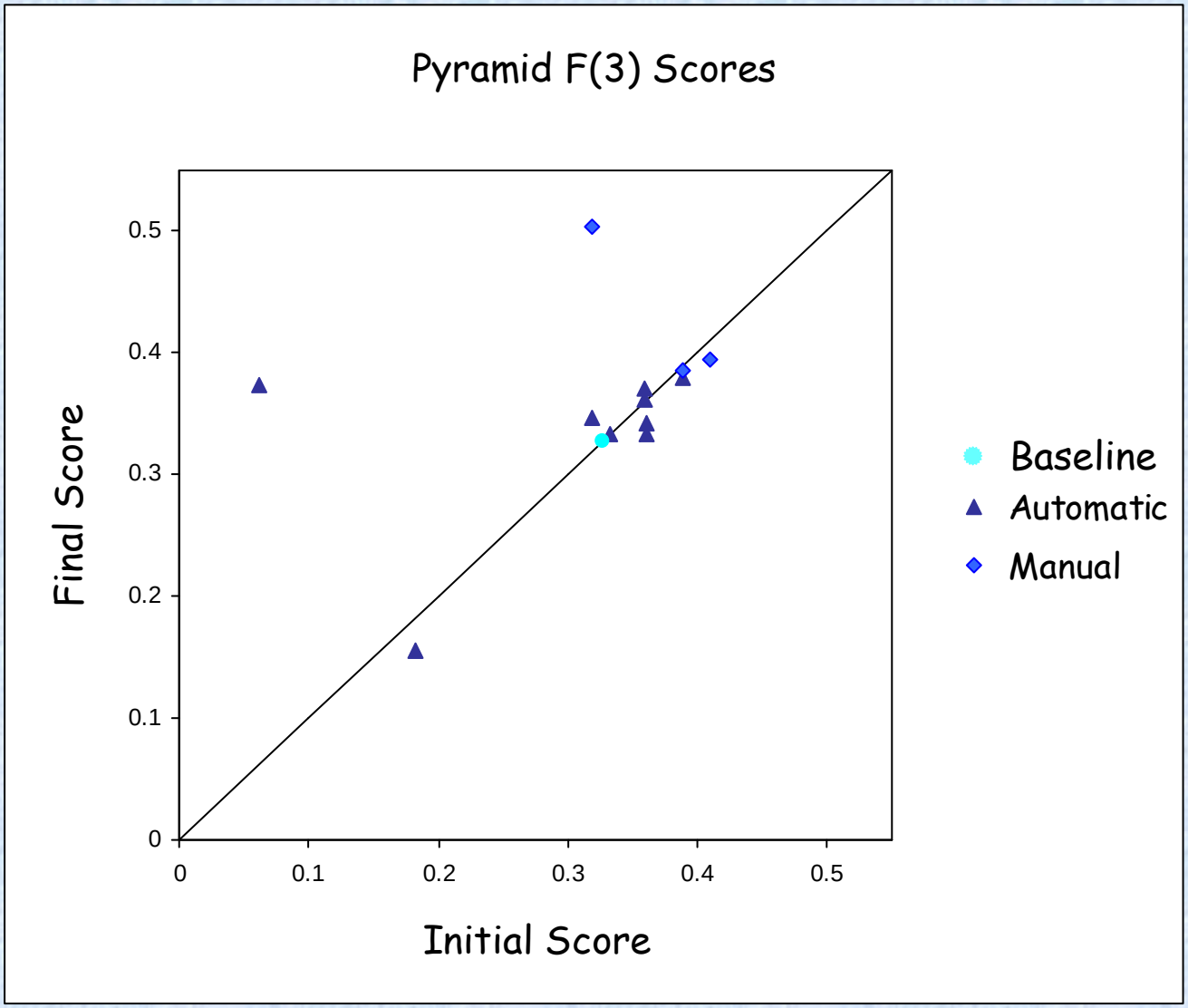
- Questions taken from relationship type identified in AQUAINT pilot
  - question formed from a relationship template
  - also included narrative giving more details

What evidence is there for transport of [goods] from [entity] to [entity]?
What [financial relationship] exists between [entity] and [entity]?
What [organizational ties] exist between [entity] and [entity]?
What [familial ties] exist between [entity] and [entity]?
What [common interests] exist between [entity] and [entity]?
What influence/effect does [entity] have on/in [entity]?
What is the position of [entity] with respect to [issue]?
Is there evidence to support the involvement of [entity] in [event/entity]?

ciQA Protocol

- Perform baseline runs
  - AQUAINT-2 corpus only
- Receive interaction responses
  - interaction via web application that asks assessor for more information
  - application designed, hosted by participant
  - assessor spends ≤ 5 minutes/topic (up from 3)
- Perform additional (non-baseline) runs exploiting additional info

ciQA Task Results



ciQA Discussion

- Exit questionnaire
  - assessors felt considerable time pressure
  - did not like "complicated" interactions, possibly because of that time pressure
- Assessors as surrogate users
  - assessors not good models for novice users
  - after creating/judging the question, the assessor is not a good model of initial information seeking behavior

Genomics Track

- Track motivation: explore information use within a specific domain
  - focus on person experienced in the domain
- 2007 task
  - similar to 2006 task: instance finding (focused response) in full text of scientific articles
  - also, compare relative effectiveness of different granularities of response

Genomics Track Task

- Documents
  - full-text journal articles provided through Highwire Press
  - associated metadata (eg MEDLINE record) available
  - 162,259 articles from 49 journals; about 12.3GB HTML
- Topics
  - 36 questions asking for lists of entities derived from interviews with working biologists
  - 13 entity types (drugs, genes, toxicities…)
    - e.g., *Which* [PATHWAYS] *are mediated by CD44?*
      What [GENES] *regulate puberty in humans?*
- System response
  - ranked list of up to 1000 passages (pieces of paragraphs)
  - each passage must contain at least one entity of target type

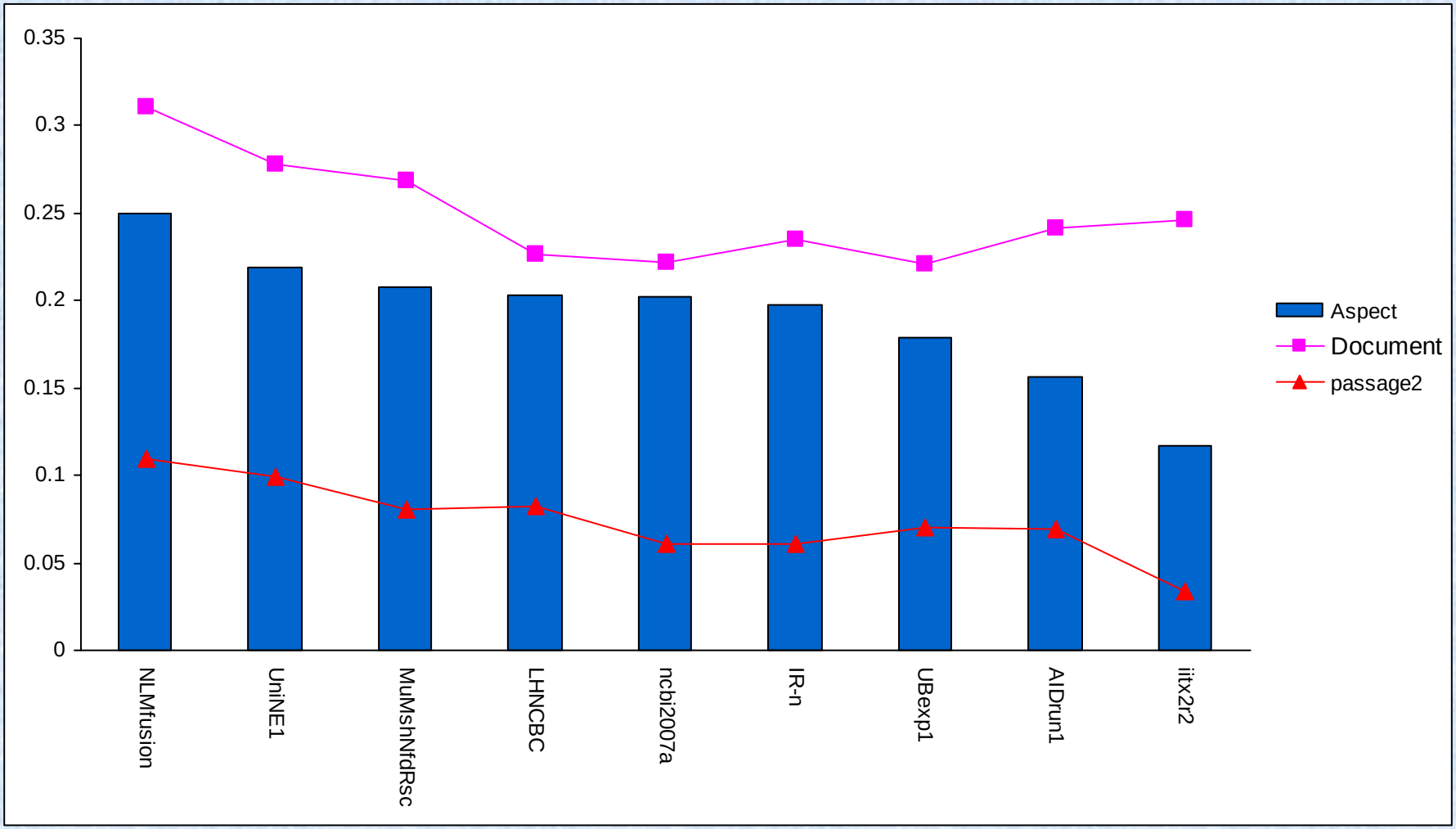
Task Evaluation

- Relevance judging
  - pools built from standardized paragraphs by mapping retrieved passage to its unique standard
  - judged by domain experts using 3-way judgments: not/possibly/definitely relevant
  - assessor marked contiguous span in paragraph as answer
- Entities
  - from set of relevant passages, assessor created gold standard list of entities
  - entities assigned to individual passages

Task Evaluation

- Scoring
  - document:
    - standard ad hoc retrieval task (MAP)
    - doc is relevant iff it contains a relevant passage
    - collapse system ranking so doc appears just once
  - aspect:
    - retrieved passage that overlaps with marked answer assigned aspect(s) of that answer; else no aspect
    - collapse ranking so instance occurs at most once
    - calculate MAP of induced ranking
  - passage ("passage2")
    - treat each character as relevant/irrelevant
    - compute MAP of character ranking

Top Automatic Runs



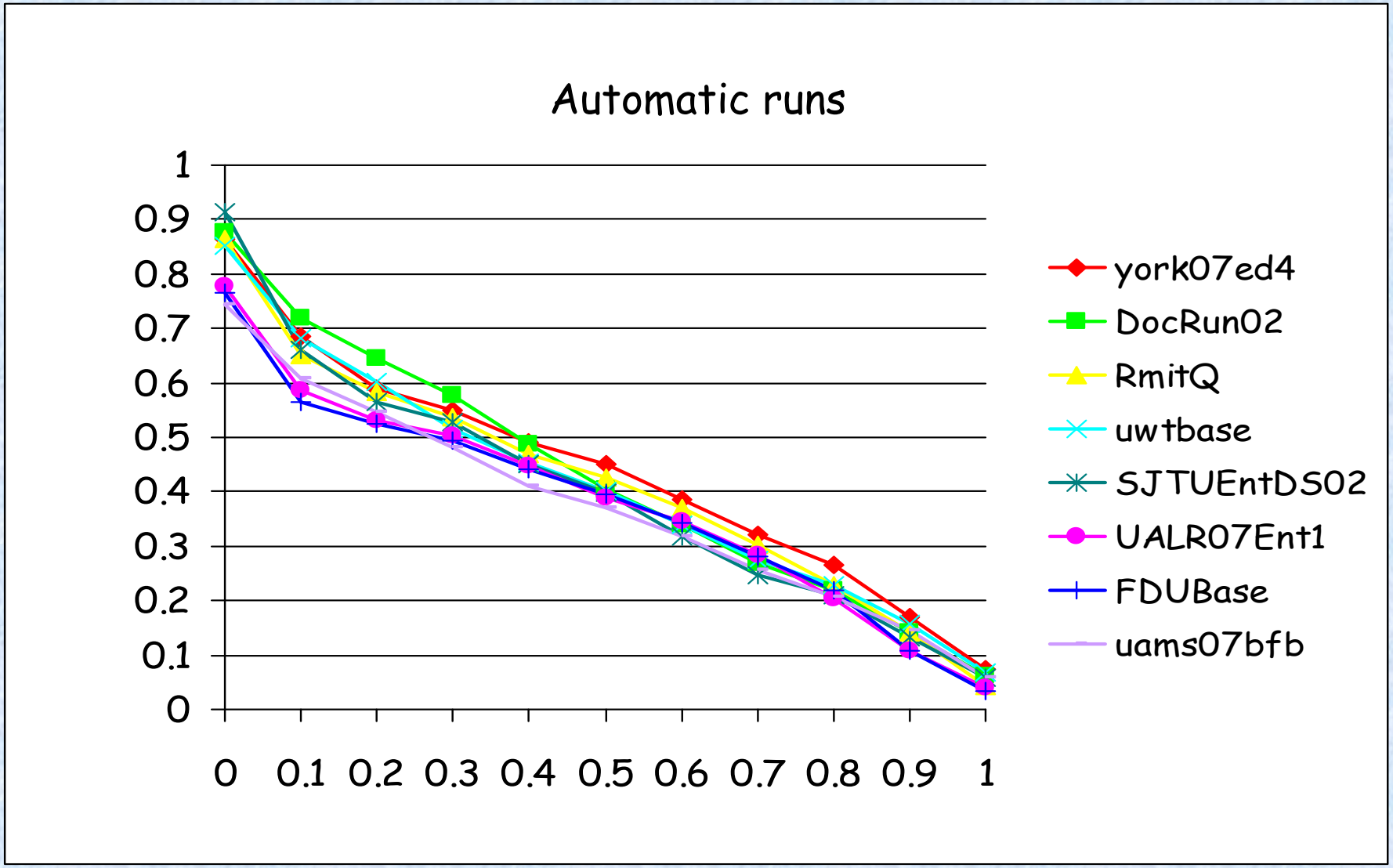
Enterprise Track

- Goal: investigate enterprise search, searching the data of an organization to complete some task
- new corpus/task for 2007
  - CSIRO science communicators target users
  - document set a crawl of .csiro.au; ~370,000 documents & 4.2 GB
  - abstract task the "missing page" problem
    - document search
    - find-an-expert task

Test Collection

- 50 topics created by real science communicators
  - query, narrative
  - some example key pages
  - list of key contacts
- Same topic set used for both tasks
  - systems given query and narrative
  - example key pages used for relevance feedback
  - key contacts list is judgments for expert search
  - community judging for document search task

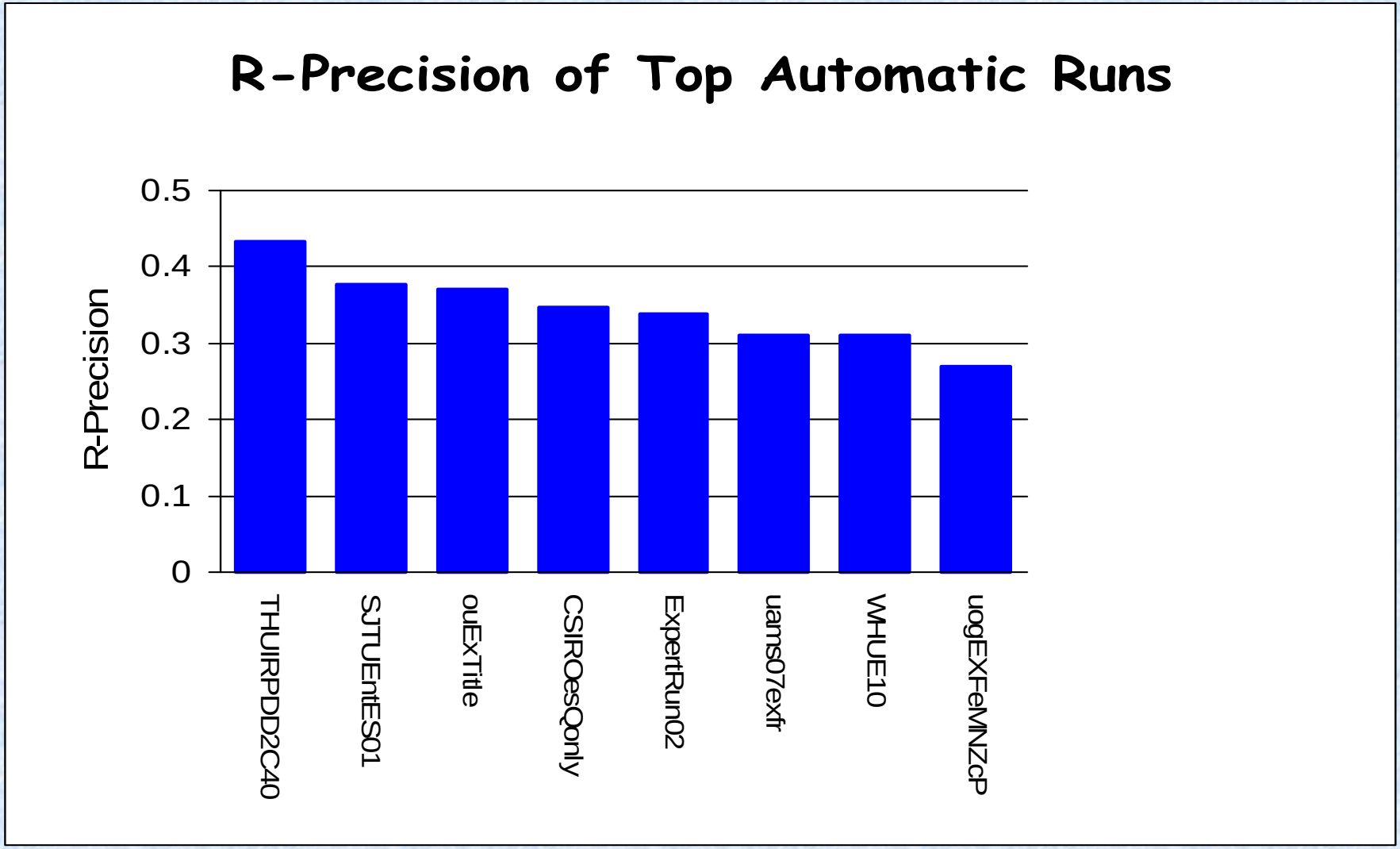
Document Search



Search-for-Experts Task

- Return a ranked list of people who are experts in the area of the topic
- People are represented by email addresses
- No list of people provided; systems extracted email address from corpus
- Evaluate as standard ad hoc retrieval task

Search-for-Experts



Blog Track

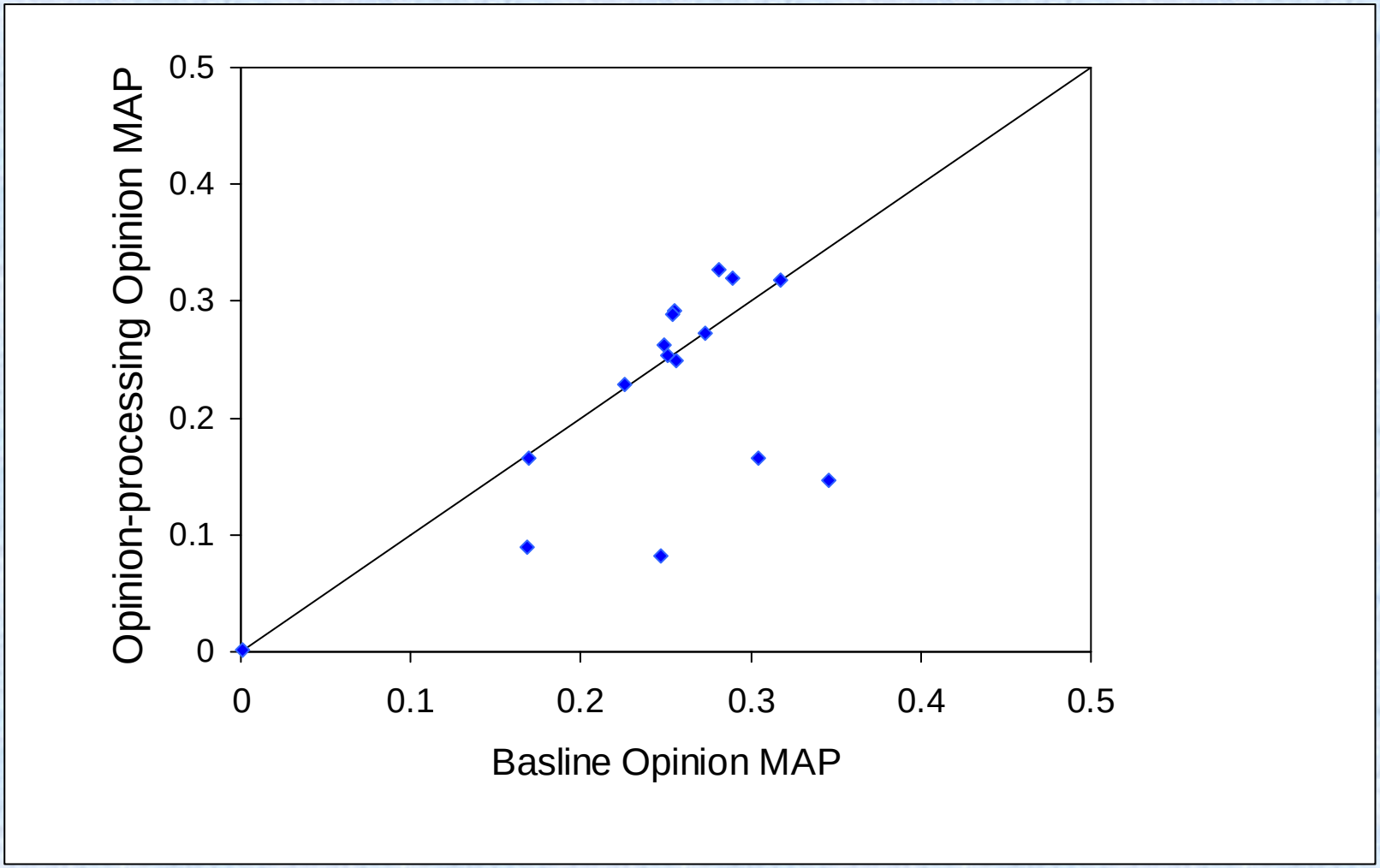
- New track in 2006
  - explore information access in the blogosphere
- Document set
  - set of blogs collected in Dec 2005-February 2006 & distributed by University of Glasgow
  - collection has 3 main components & miscellaneous such as spam, RSS feeds, non-English docs
    - 38.6GB (~100,000) feeds (i.e., different blogs)
    - 88.8GB (~3.2 million) permalinks (blog entry + comments)
    - 28.8GB home pages

Blog Tasks

- Opinion task
  - find posts that express an opinion about target (person, organization, brand, technology, etc)
  - document is a permalink
  - polarity subtask: classify opinion as positive, negative, mixed
  - topics, relevance judgments by NIST assessors
- Blog distillation (feed search)
  - find blog with primary, recurring interest in topic
  - document is feed (blog as a whole)
  - topics, relevance judgments by community

Opinion Task

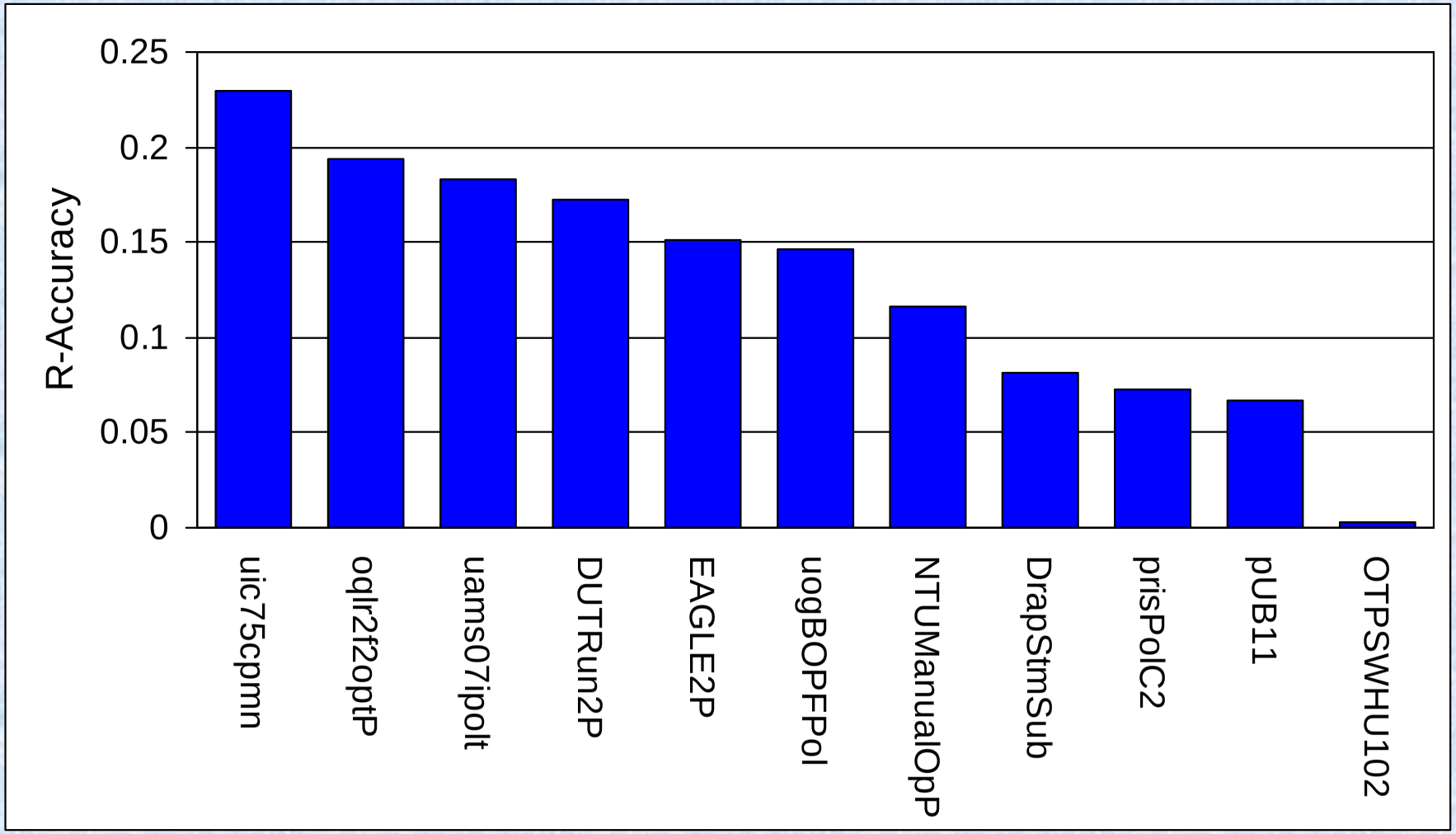
Effect of adding opinion-specific processing to baseline runs;
Best title-only, automatic run pair per group



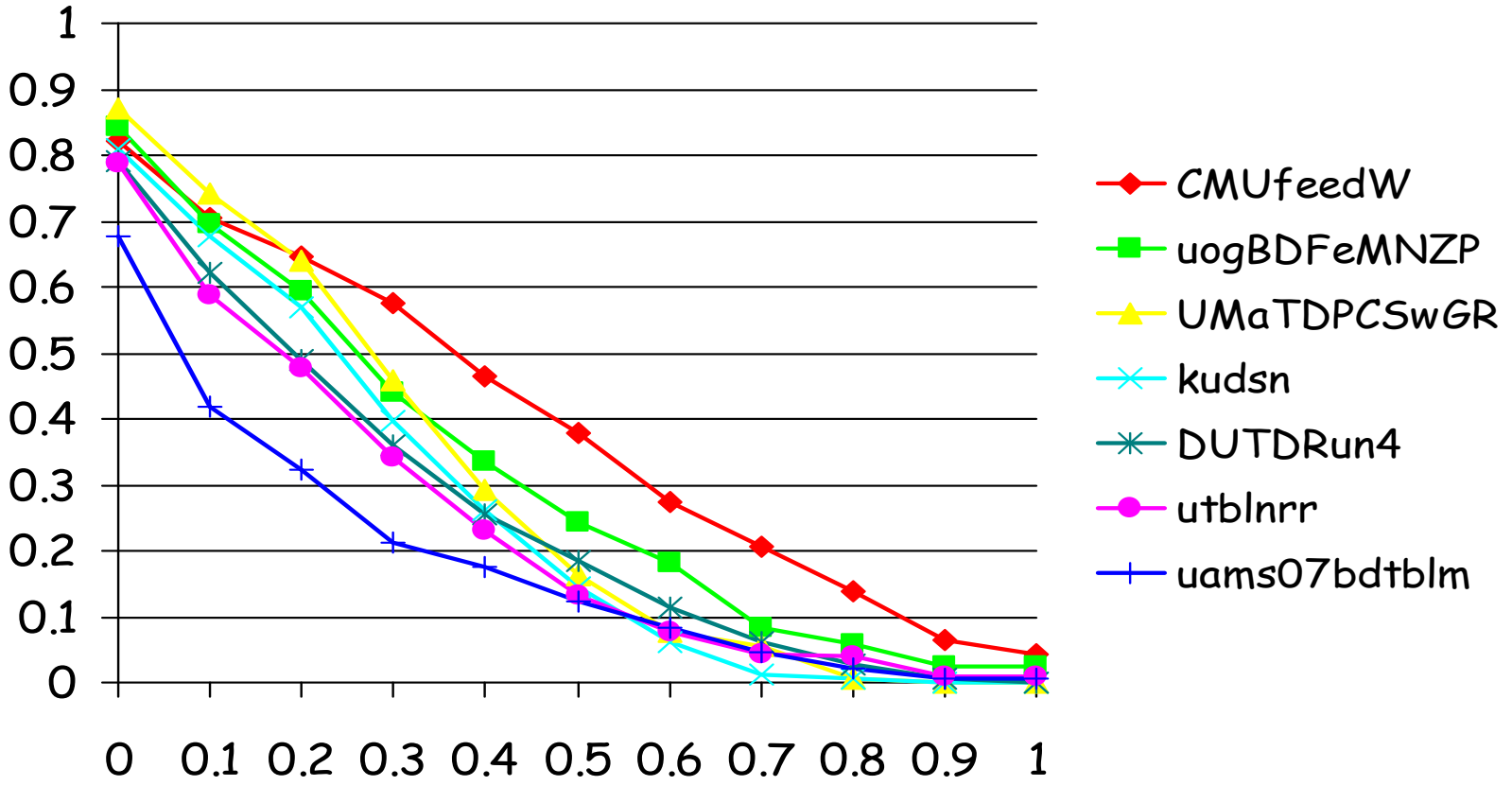
Polarity Subtask

- For same ranking as submitted to opinion task, label the documents
  - labels same as judgments: positive, negative, mixed
  - if opinionated but unclear how, tagged as mixed
- Evaluation
  - not typical classification problem since not all documents classified
  - different systems classified different documents
  - use "R-accuracy": R-Precision where document is treated as relevant only if correctly labeled

Polarity Results



Feed Distillation



Legal Track

- Goal: evaluate search technology for discovery of electronically stored information in litigation and regulatory settings
  - domain has a high-recall focus
- First run in 2006, coinciding with changes to the Federal Rules of Civil Procedure
  - electronically stored information (ESI) on par with all other "documents"

Legal Track Collection

- Document set
  - almost 7 million documents (scientific reports, memos, email, budgets…) made public through the tobacco Master Settlement Agreement
  - IIT CDIP Test Collection, version 1.0 (OCR)
- Topics
  - 4 hypothetical complaints
  - 50 requests to produce (43 judged)
  - topic statement included Boolean queries and B, size of the retrieved set of final Boolean query
  - negotiated in accordance with standard practice by lawyers

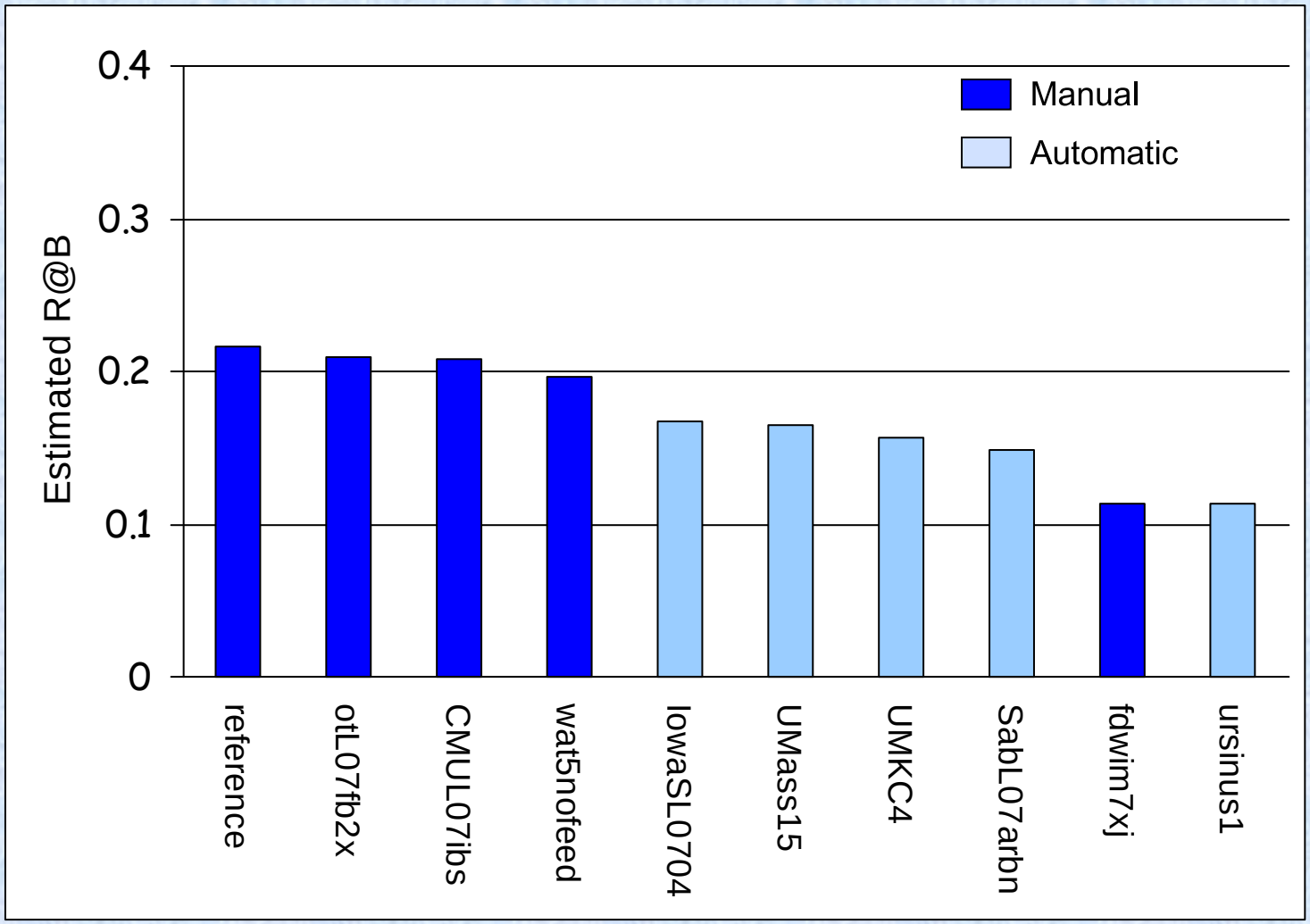
Legal Track Collection

- Relevance judgments
  - specific goal of measuring recall at B required different pooling strategy
    - large, messy collection meant standard pooling would need too large (expensive) a cut-off level
    - Boolean output unranked
  - all submitted runs (plus a "random" run) sampled using strategy that gives best estimate for recall at B given maximum pool size
  - unclear to what extent collection is reusable for other runs, other measures
  - sampling/evaluation done by Stephen Tomlinson
  - relevance judging done (mostly) by law students

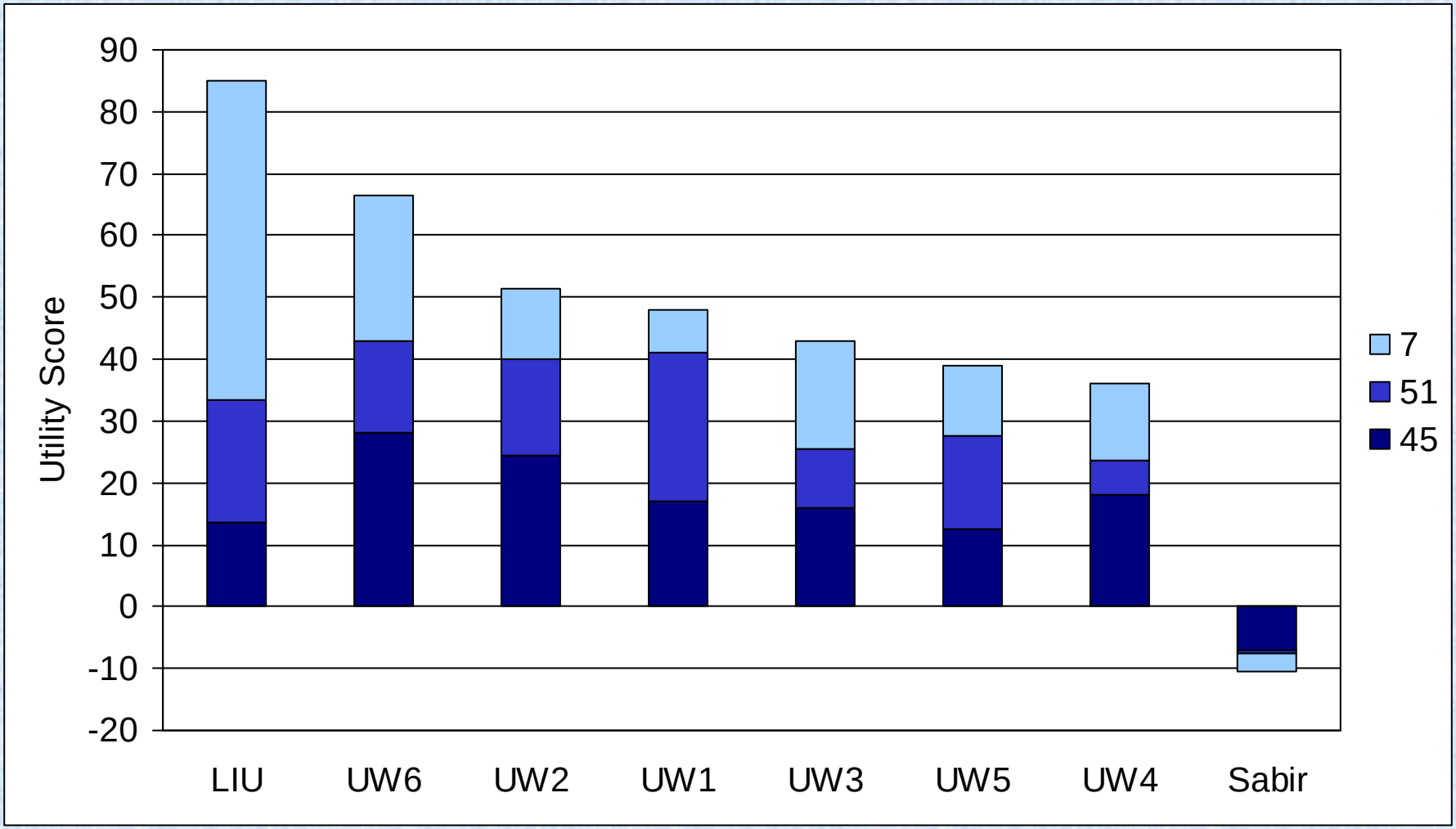
Legal Track Tasks

- Ad hoc
  - same as 2006 task modulo a few tweaks
  - new topics against document collection
- Interactive
  - humans use whatever search protocol desired to find up to 100 new relevant docs for 2006 topics
  - evaluate by utility that penalizes retrieved nonrel
- Relevance Feedback
  - automatic retrieval using 2006 topics plus relevance judgments from 2006
  - same 10 topics as used in interactive assessed

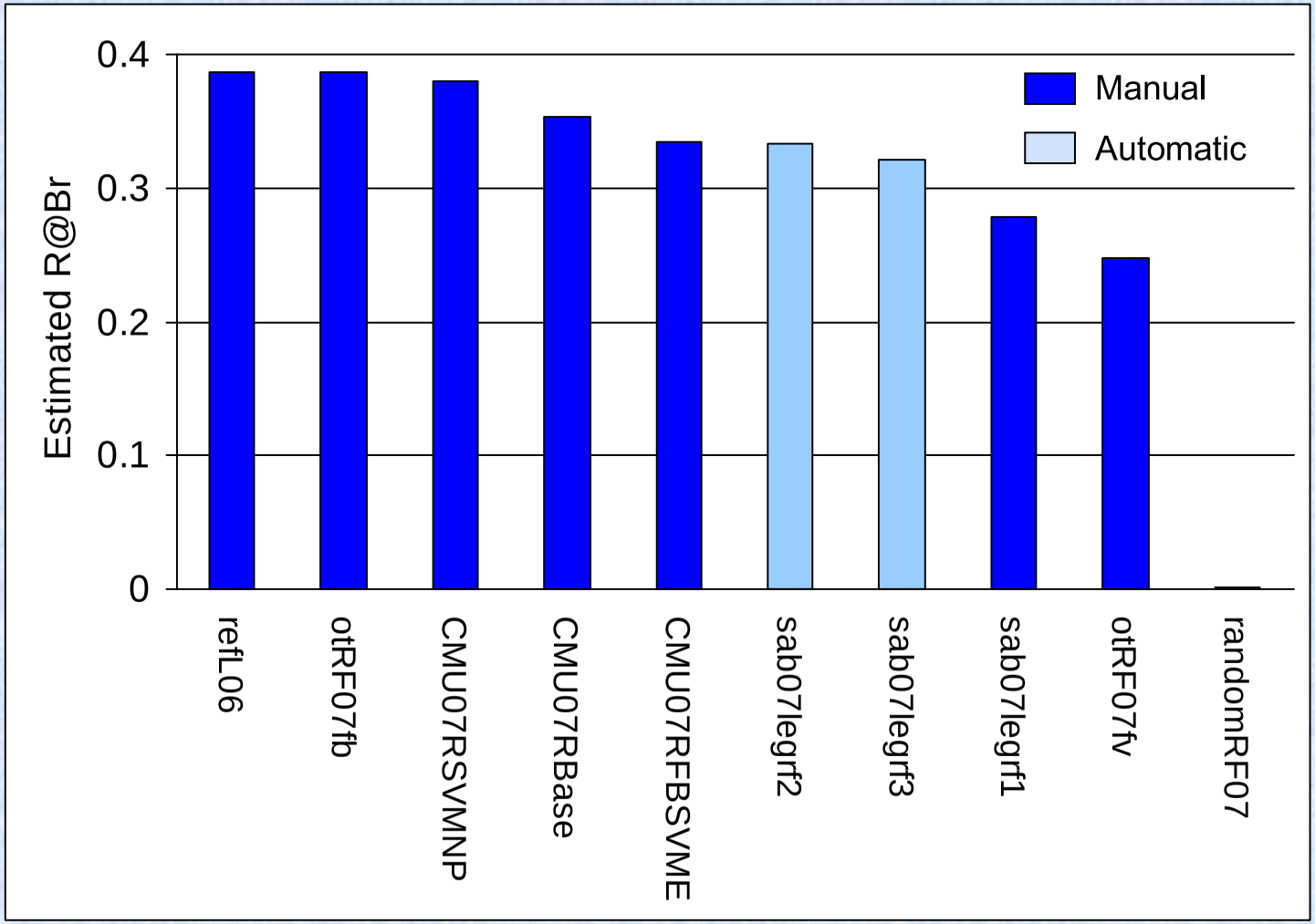
Ad Hoc Results



Interactive Results



Feedback Results



Spam Track

- Motivation:
  - assess quality of an email spam filter's actual usage
  - lay groundwork for other tasks with sensitive data
- How to get appropriate corpus?
  - true mail streams have privacy issues
  - simulated/cleansed mail streams introduce artifacts that affect filter performance
  - track solution: create software jig that applies given filter to given message stream and evaluates performance based on judgments
  - have participants send filters to data

Spam Tasks

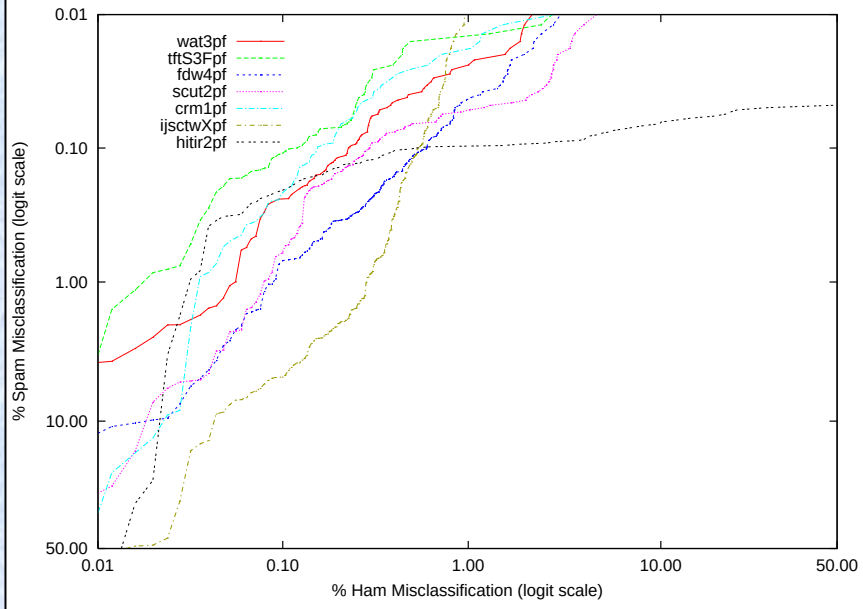
- 2 email streams [ham; spam; total]
  - trec07p (public) [25,220; 50,199; 75,419]
    - all messages delivered to a server Apr 8–Jul 6
  - MrX3 (private) [8,082; 153,893; 161,975]
    - all of X's email from Dec 2006---July 11, 2007
    - big increase in spam, constant ham across 3 years
- 4 tasks
    - immediate feedback filtering
    - delayed feedback filtering
    - partial feedback filtering (public only)
    - active learning

Evaluation

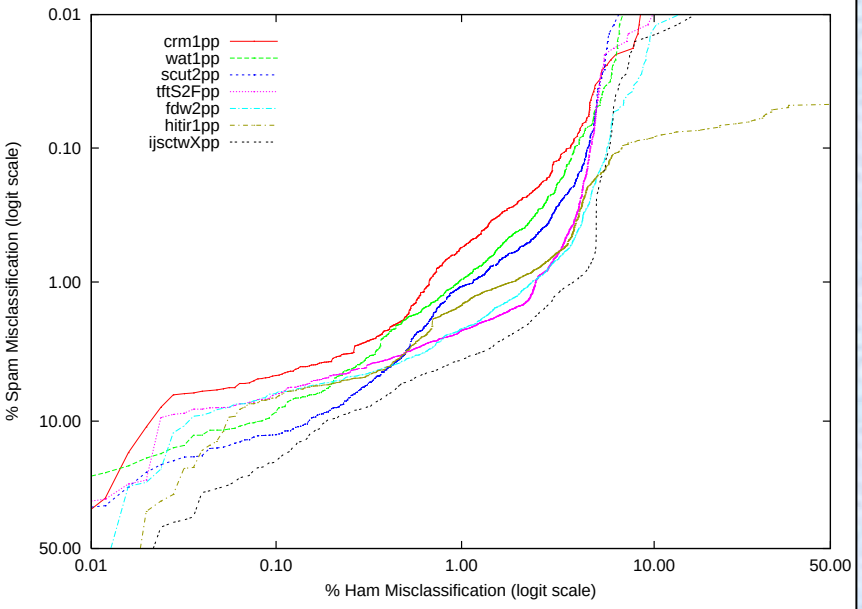
- Ham misclassification rate (hm%)
- Spam misclassification rate (sm%)
- ROC curve
  - assumes filter computes a "spamminess" score
  - use score to compute sm% as function of hm%
  - area under ROC curve is measure of filter effectiveness
  - use 1-area expressed as a % to reflect filter ineffectiveness (1-ROCA)%

Immediate vs. Partial Feedback

trec07p corpus



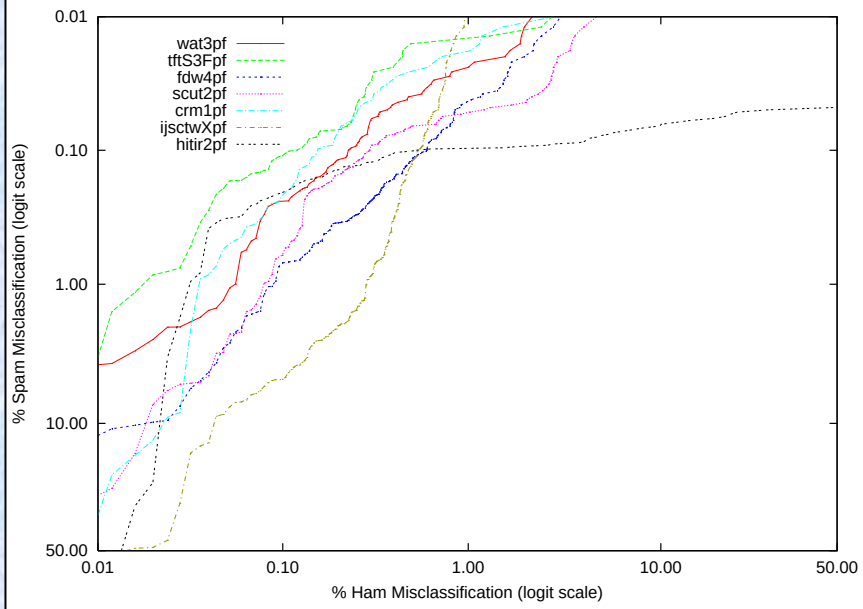
Immediate Feedback



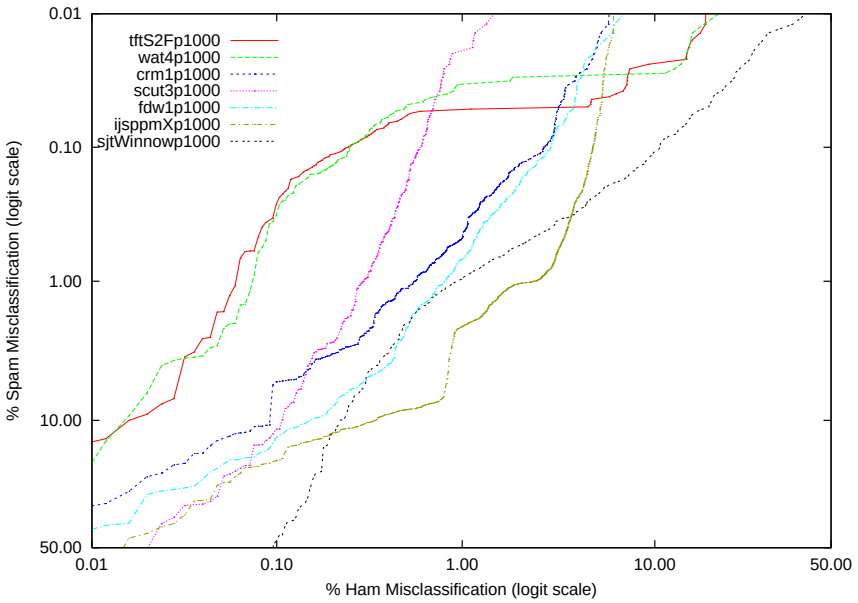
Partial Feedback

Immediate Feedback vs. Active Learning

trec07p corpus



Immediate Feedback



Active Learning (quota 1000)

Future

- TREC expected to continue into 2008
- TREC 2008 tracks:
  - sunset genomics, spam
  - move QA to TAC (expanded of what was DUC)
  - blog, enterprise, legal, million query tracks continue
  - add relevance feedback track
    - goal: create evaluation methodology/data that will allow separating effects of different variables so as to improve effectiveness
    - use Q0 field in TREC submission format ✪
  - track on utility of tags under development, probably can't get data in time for 2008

Track Brainstorming

- Thursday, 10:10 Lecture Room A (during break)
  - solicit ideas for what future TRECs should be concerned with
  - true brainstorming: "wild" ideas encouraged; no filtering by resource requirements
  - might be possible to incorporate into TREC 2008; more likely, further out
  - relatively early in conference to facilitate further discussion; also feel free to contact PC members